AD-A033 602 ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF MATHEMATICS F/G 12/1
EXAMINATION AND ANALYSIS OF RESIDUALS: A TEST FOR DETECTING A M--ETC(U)
NOV 76 A HEDAYAT, B L RAKTOE , P P TALWAR AF-AFOSR-3050-76
UNCLASSIFIED AFOSR-TR-76-1256 NL

1 OF 1
AD
A033602

END
DATE
FILMED
2-77

AFOSR - TR - 76 - 1256

3

BS

See 1473

3

EXAMINATION AND ANALYSIS OF RESIDUALS: A TEST FOR DETECTING A MONOTONIC RELATION BETWEEN MEAN AND VARIANCE IN REGRESSION THROUGH THE ORIGIN

by

A. Hedayat[1], B.L. Raktoe[2] and Prem P. Talwar[2]

Department of Mathematics
University of Illinois at Chicago Circle[1]
University of Guelph[2]

November 9, 1976

PREPARED UNDER GRANT AFOSR 76-3050 ✓
FOR
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ILLINOIS AT CHICAGO CIRCLE
CHICAGO, ILLINOIS 60680

EXAMINATION AND ANALYSIS OF RESIDUALS: A TEST FOR DETECTING A MONOTONIC RELATICN BETWEEN MEAN AND VARIANCE IN REGRESSION THROUGH THE ORIGIN

A. Hedayat[1], B. L. Raktoe[2] and Prem P. Talwar[2]

[1]University of Illinois at Chicago Circle

[2]University of Guelph

ABSTRACT

This paper presents a simple and exact test for detecting a monotonic relation between the mean and variance in linear regression through the origin. This test resulted from utilizing uncorrelated Theil-residuals and the Goldfeld-Quandt peak test. A numerical example is provided to elucidate the method. A simulation experiment was performed to compare the empirical power of this test with those of the existing tests.

1. INTRODUCTION

Consider the simple linear model $Y = X\beta + \varepsilon$, where Y is an n-dimensional random vector of observations, X is an n-dimensional vector consisting of known nonstochastic elements, β is an unknown scalar and ε is an n-dimensional random vector, and

$$E[\varepsilon] = 0 \ , \quad E[\varepsilon \varepsilon'] = \sigma^2 I_n \ , \qquad (1.1)$$

where $\sigma^2 > 0$ is an unknown parameter and I_n is the $n \times n$ identity matrix.

The least squares (LS) estimator $\hat{\beta}$ of β and the least squares predictor $\hat{\varepsilon}$ of ε are given by

$$\hat{\beta} = (\sum_{i=1}^{n} x_i\, y_i)(\sum_{i=1}^{n} x_i^2)^{-1} \text{ , and}$$

$$\hat{\varepsilon} = Y - X\hat{\beta} = Y - X\,(\sum_{i=1}^{n} x_i\, y_i)(\sum_{i=1}^{n} x_i^2)^{-1} = PY \text{ ,}$$

where

$$P = I_n - (\sum_{i=1}^{n} x_i^2)^{-1}\, XX' \text{ .}$$

Under the assumptions (1.1)

$$E[\hat{\varepsilon}] = PX\beta = 0\beta = 0$$

$$E[\hat{\varepsilon}\ \hat{\varepsilon}'] = P\ E[YY']P' = P\ E[\varepsilon\ \varepsilon']P' = \sigma^2 P \text{ .}$$

Hence it is clear that even when (1.1) holds, the LS estimators of residuals are neither independent nor do they have constant variance since $P \neq I_n$.

Goldfeld and Quandt [1965] present two exact tests for testing the hypothesis that the residuals from a least squares regression are homoscedastic. The first test is parametric and uses the F-statistic. The second test is nonparametric and uses the number of peaks in the ordered sequence of unsigned residuals. Hedayat and Robson [1970], among other results, have demonstrated the failure of Goldfeld and Quandt peak test applied to LS residuals. One reason of the failure is that least squares residuals, even under ideal conditions, are in general correlated and have different variances.

In this paper, we work with a different type of residuals which are free from the above criticism. We will use the new residuals to detect a monotonic relation between the mean and variance by means of the peak test introduced by Goldfeld and Quandt [1965].

2. T-RESIDUALS AND THEIR PROPERTIES IN SIMPLE LINEAR REGRESSION THROUGH THE ORIGIN

Theil [1965] has presented a predictor of ε (designated by T-residuals) which has all the ordinary properties of $\hat{\varepsilon}$, except that the covariance matrix of T-residuals is $\sigma^2 I_{n-1}$ under the assumption (1.1). Koerts [1967] derived the explicit form of the T-residuals for the simple linear model through the origin. Following Koerts the elements of the vector of T-residuals ε^* can be represented by

$$\varepsilon^*_i = y_i - b^* x_i \ , \quad i = 1, 2, \ldots, n \ , \quad i \neq k \ ,$$

where:

$$b^* = \left[1 - |x_k| \left(\sum_{i=1}^{n} x_i^2 \right)^{-\frac{1}{2}} \right] \hat{\beta}_{n-1} + \left[|x_k| \left(\sum_{i=1}^{n} x_i^2 \right)^{-\frac{1}{2}} \right] y_k \, x_k^{-1} \ ,$$

and

$$\hat{\beta}_{n-1} = \left(\sum_{\substack{i=1 \\ i \neq k}}^{n} x_i \, y_i \right) \left(\sum_{\substack{i=1 \\ i \neq k}}^{n} x_i^2 \right)^{-1} .$$

In the above expression k can take any value from 1 to n. The properties of T-residuals are the following:

(i) ε^*_i is a linear function of y_i ,

(ii) $E[\varepsilon^*_i] = 0 \ , \quad i = 1, 2, \ldots, n \ , \quad i \neq k$,

(iii) $$\mathrm{Cov}[\varepsilon^*_i, \varepsilon^*_j] = \begin{cases} 0 \ , & \text{if } i \neq j \\ \sigma^2 \ , & \text{if } i = j \end{cases}$$

where $i, j = 1, 2, \ldots, n \ , \quad i, j \neq k$,

(iv) The T-residuals have a minimum expected sum of squares of errors $(\varepsilon_i^* - \varepsilon_i)$ in the class of predictors satisfying properties (i), (ii) and (iii), and

(v) $\sum_{\substack{i=1 \\ i \neq k}}^{n} \varepsilon_i^{*2} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$.

As can be seen and in light of the remarks we made earlier, properties (iii) and (v) make the T-residuals very interesting indeed. T-residuals have been derived based on the first four properties and Koerts [1967] has shown that they also have the fifth property.

3. A SIMPLE AND EXACT TEST WHICH DETECTS MONOTONICITY OF VARIANCES IN SIMPLE LINEAR REGRESSION THROUGH THE ORIGIN

Consider the case where the x_i's have been ordered such that $x_i < x_j$ if $i < j$ and suppose our interest lies in testing the following hypothesis:

$$\begin{aligned} H_0&: \; E[\varepsilon_i^2] = \sigma^2 \quad \text{against} \\ H_1&: \; E[\varepsilon_i^2] = \sigma_i^2 < E[\varepsilon_j^2] = \sigma_j^2 \quad \text{for} \quad i < j \, . \end{aligned} \tag{3.1}$$

Note that the alternative hypothesis says that as x increases the variance of ε or y also increases. We are considering the case where we have only a single observation for each level x , as is frequently the case.

Two alternative tests for testing H_0 against H_1 are suggested by Goldfeld and Quandt [1965], namely:

(i) The F test

The obvious choice for k is then the middle observation, so that one can compute the ratio of the sum of squares of the first $(n - 1)/2$ predicted residuals to that of the last $(n - 1)/2$, which is F distributed. When $n - 1$ is not even, one can use either $(n - 2)/2$ first and $n/2$ last observations or $n/2$ first and $(n - 2)/2$ last observations, and for this choice see Theil [1965].

(ii) The Peak test

For residuals ordered by the ordering of x_i, $x_i < x_{i+1}$, define a peak at x_i to be an instance where $|\hat{\varepsilon}_i| > |\hat{\varepsilon}_j|$ for $j = 1, 2, \ldots, i - 1$.

The validity of applying the Goldfeld Quandt peak test to the T-residuals is seen by noting that under H_0, the ε_i^*'s are uncorrelated so that under the normality assumption they will be independent.

In the class of regressions restricted by the conditions that the x_i's are positive and

$$\sigma_i^2/\sigma_j^2 < \frac{[(c_1x_j^2 - 1)^2 - c_1^2x_i^2x_j^2]}{[(c_1x_i^2 - 1)^2 - c_1^2x_i^2x_j^2]}$$

where c_1 is given below, we show that under H_1, $\text{var}[\varepsilon_i^*] < \text{var}[\varepsilon_j^*]$. This means that especially in such settings a greater sensitivity can be expected of the peak test based on the T-residuals than from the F-test, which is a general test.

THEOREM 3.1. *If* $E[\varepsilon_i \varepsilon_j] = 0$, $i \neq j$, *and* $E[\varepsilon_i^2] = \sigma_i^2 < E[\varepsilon_j^2] = \sigma_j^2$, *then* $\text{var}[\varepsilon_i^*] < \text{var}[\varepsilon_j^*]$ *if* $x_t > 0$, $\forall\, t$ *and*

$$\sigma_i^2/\sigma_j^2 < \frac{[(c_1x_j^2 - 1)^2 - c_1^2x_i^2x_j^2]}{[(c_1x_i^2 - 1)^2 - c_1^2x_i^2x_j^2]} .$$

Proof. Under these assumptions and by definition of ε_i^*

$$\text{var}[\varepsilon_i^*] = E[\varepsilon_i^{*2}] - (E[\varepsilon_i^*])^2 = E[\varepsilon_i^{*2}]$$

$$= \sigma_i^2(c_1x_i^2 - 1)^2 + c_3c_1^2x_i^2 + c_1^2x_i^2x_j^2\sigma_j^2 + (\sigma_k^2x_i^2)/\sum_{t=1}^{n} x_t^2 ,$$

where

$$c_1 = \left[1 - |x_k| \left(\sum_{t=1}^{n} x_t^2 \right)^{-\frac{1}{2}} \right] \left(\sum_{t \neq k}^{n} x_t^2 \right)^{-1}$$

$$c_2 = \sigma_k^2 \left(\sum_{t=1}^{n} x_t^2 \right)^{-1} \quad \text{and}$$

$$c_3 = \sum_{t \neq i,j,k} x_t^2 \, \sigma_t^2 \quad .$$

$$\text{var}[\epsilon_j^*] - \text{var}[\epsilon_i^*] = \sigma_j^2 \, (c_1 x_j^2 - 1)^2 - \sigma_i^2 \, (c_1 x_i^2 - 1)^2$$
$$+ \, c_3 c_1^2 \, (x_j^2 - x_i^2) + c_1^2 x_i^2 x_j^2 \sigma_i^2 - c_1^2 x_i^2 x_j^2 \sigma_j^2 + c_2 \, (x_j^2 - x_i^2) \; .$$

Since $x_i < x_j$ and they are positive, it follows that in order to show $\text{var}[\epsilon_j^*] - \text{var}[\epsilon_i^*] \geq 0$ it is sufficient to show that

$$\sigma_j^2 \, (c_1 x_j^2 - 1)^2 - \sigma_i^2 \, (c_1 x_i^2 - 1)^2 + c_1^2 x_i^2 x_j^2 \sigma_i^2 - c_1^2 x_i^2 x_j^2 \sigma_j^2 \geq 0$$

and this will be true if and only if

$$\sigma_i^2 / \sigma_j^2 < \frac{[(c_1 x_j^2 - 1)^2 - c_1^2 x_i^2 x_j^2]}{[(c_1 x_i^2 - 1)^2 - c_1^2 x_i^2 x_j^2]} \quad .$$

4. A NUMERICAL ILLUSTRATION

To elucidate the use of our peak test we go for the benefit of the reader through a complete example. Let us consider the example (see Table I) given on page 180 of Steel and Torrie [1960]. As these authors have pointed out, in this instance the regression line should pass through the origin. Therefore, $\hat{\beta} = 3.67$ and hence the regression line is given by $y = 3.67x$. The individual least square residuals, after rounding to one decimal place, are given in Table I.

TABLE I

Induced reversions to independence per 10^7 surviving cells y per dose (ergs/Bacterium) $10^{-5}x$ of Streptomycin dependent Escherichia Coli subjected to monschromatic ultraviolet radiation of 2,967 Angstroms wave length.

x	y	$\hat{\varepsilon}$
13.6	52	2.0
13.9	48	-3.1
21.1	72	-5.5
25.6	89	-5.1
26.4	80	-17.0
39.8	130	-16.2
40.1	139	-8.3
43.9	173	11.7
51.9	208	17.3
53.2	225	29.5
65.2	259	19.5
66.4	199	-45.0
67.7	255	6.3

First of all, visual examination of these residuals suggests, that there is a pattern for the distribution of plus and minus signs among the $\hat{\varepsilon}_i$'s . Secondly, graphical plotting of residuals against the fitted values or x-values strongly suggests that the error variance increases with x . Now, suppose we suspect the assumption $E[\varepsilon_1^2] = \sigma^2$ for all i and in particular we suspect that the variance may increase with the mean, i.e. that the variance of y increases as x increases. To test against this alternative hypothesis we first compute the T-residuals. We note that under H_0

the distribution of the number of peaks is independent of the choice of k, which depends primarily on the power of the test with respect to a specific alternative hypothesis. However, it seems that the index of the middle observation would be a reasonable choice of k for our general H_1. Recall that H_1 puts no restriction on the monotonicty structure of the variance other than being increasing. If we let $k = 7$, we have

$$\varepsilon_i^* = y_i - b^* x_i \ , \quad i = 1, 2, \ldots, 6, 8, \ldots, 13 \tag{4.1}$$

where $b^* = 3.63$. Thus, the individual T-residuals, after rounding to one decimal place, are as follows:

$\varepsilon_1^* = +2.6$	$\varepsilon_8^* = +13.5$
$\varepsilon_2^* = -2.5$	$\varepsilon_9^* = +19.4$
$\varepsilon_3^* = -4.6$	$\varepsilon_{10}^* = 31.7$
$\varepsilon_4^* = -4.0$	$\varepsilon_{11}^* = 22.1$
$\varepsilon_5^* = -15.9$	$\varepsilon_{12}^* = -42.2$
$\varepsilon_6^* = -14.6$	$\varepsilon_{13}^* = +9.0$

The number of peaks is 5.

The ε_i^*'s are independent and identically distributed under the homoscedasticity and normality assumptions of the ε_i's. Now, we can compute the probability of obtaining five or more peaks in a sequence of 12 independent and identically distributed random variables using Table I from Goldfeld and Quandt [1965]. By interpolation from this table we see that this probability is about .036. If we can accept a risk of 3.6 percent, then we should fit a weighted regression rather than the unweighted one for obtaining an efficient estimate of β and hence the regression line.

5. SIMULATION STUDY

We consider the simple model $y_i = x_i\ (\beta + \varepsilon_i)$,

$i = 1, 2, \ldots, n$. Sampling experiments were performed on this model in order to obtain empirical estimates of the powers of three tests 1) F-test, 2) Goldfeld-Quandt **peak** test and 3) Peak test based on the uncorrelated T-residuals. The independent variable was identical in repeated samples and each particular sample of x's was chosen from the uniform distribution with mean $\mu_x = 30, 40, 50$ and standard deviation $\sigma_x = 10, 20, 25$.. The total number of observations was 31 . For each μ_x , σ_x combination, one sample of x's was generated and for each such sample, 1000 samples of 31 ε-values were generated. In our simulation study we considered three distributions for the errors ε

a) the normal distribution with zero mean and unit variance

b) the student's "t" with 2 degrees of freedom (d.f.) and

c) the adjusted chi-square distribution with 4 d.f., adjusted so that the mean is equal to zero.

Uniform pseudorandom numbers were generated by a multiplicative-congruential method of an IBM 360/65. The uniform variates were used to form observations from the distribution studied; the Gaussian by a modification of the Box-Muller method; the chi-square with 4 d.f. as -2 times the logarithm of the product of 2 indĕpendent uniform random numbers; and the t with 2 d.f. as the ratio of a Gaussian and the square root of a chi-square with 2 d.f.

The Monte Carlo results for the various distributions are given in Table II. The simulation results clearly establish the superiority of the peak-test based on T-residuals over the other two tests in case of normal and chi-square distributions. In case of "t" with 2 d.f., F-test compare favorably with Peak test on T-residuals.

TABLE II

Empirical Power for Nominal Size of .05

a) Distribution of errors: normal, mean = 0, variance = 1 .

μ_X	σ_X	F-test	Peak Test on LS Residuals	Peak Test on T-Residuals
30	10	.045	.015	.144
	20	.023	.008	.751
	25	.018	.006	.521
40	10	.052	.020	.08
	20	.031	.015	.339
	25	.025	.007	.669
50	10	.052	.02	.059
	20	.039	.016	.209
	25	.031	.015	.339

b) Distribution of errors: t with 2 d.f.

30	10	.206	.012	.081
	20	.156	.011	.426
	25	.128	.008	.253
40	10	.22	.013	.050
	20	.180	.012	.157
	25	.157	.015	.360
50	10	.233	.014	.037
	20	.199	.015	.106
	25	.181	.012	.157

c) Distribution of errors: adjusted chi-square with 4 d.f.

30	10	.144	.015	.13
	20	.164	.018	.617
	25	.165	.018	.402
40	10	.143	.018	.075
	20	.161	.015	.302
	25	.164	.017	.548
50	10	.136	.016	.058
	20	.149	.013	.174
	25	.161	.015	.302

BIBLIOGRAPHY

Anscombe, F. J. (1961). Examination of residuals. *Proc. Fourth Berkeley Symp. Math. Statist. I*, 1-36.

Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics* 5, 141-160.

Goldfeld, S. M. and Quandt, R. E. (1965). Some tests for homoscedasticity. *J. Amer. Statist. Assoc. 60*, 539-547.

Hedayat, A. and Robson, D. S. (1970). Independent stepwise residuals for testing homoscedasticity. *J. Amer. Statist. Assoc. 65*, 1573-1581.

Koerts, J. (1967). Some further notes on disturbance estimates in regression analysis. *J. Amer. Statist. Assoc. 62*, 169-183.

Steel, R. G. D. and Torrie, J. H. (1960). *Principle and Procedures of Statistics*. New York: McGraw Hill.

Theil, H. (1965). The analysis of disturbances in regression analysis. *J. Amer. Statist. Assoc. 60*, 1067-1079.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 18 19 AFOSR - TR - 76 - 1256

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle) 6

EXAMINATION AND ANALYSIS OF RESIDUALS: A TEST FOR DETECTING A MONOTONIC RELATION BETWEEN MEAN AND VARIANCE IN REGRESSION THROUGH THE ORIGIN.

5. TYPE OF REPORT & PERIOD COVERED 9

Interim rept.,

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s) 10

A. Hedayat, B.L. Raktoe [redacted] Prem P. Talwar

8. CONTRACT OR GRANT NUMBER(s) 15

✓ AF-AFOSR [redacted]-3050 - 76

9. PERFORMING ORGANIZATION NAME AND ADDRESS

University of Illinois at Chicago Circle
Department of Mathematics
Box 4348, Chicago, Illinois 60680

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

61102F 16 2304/A5 17 A5

11. CONTROLLING OFFICE NAME AND ADDRESS

Air Force Office of Scientific Research/NM
Bolling AFB, Washington, D.C. 20332

12. REPORT DATE 11 9

Nov [redacted] 1976

13. NUMBER OF PAGES

11 pages

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

12 14p.

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Analysis of Residuals, Regression,

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper presents a simple and exact test for detecting a monotonic relation between the mean and variance in linear regression through the origin. This test resulted from utilizing uncorrelated Theil-residuals and the Goldfeld-Quandt peak test. A numerical example is provided to elucidate the method. A simulation experiment was performed to compare the empirical power of this test with those of the existing tests.

409 950 LB